https://doi.org/10.21289/KSIC.2024.27.4.925

BERT 모델 기반 기술융합기회 탐색 연구: 웨어러블 기술사례를 중심으로

Exploration of Technology Convergence Opportunities Based on BERT Model: The Case of Wearable Technology

박진우[1], 송지훈[2*]

Jinwoo Park[1], Chie Hoon Song[2*]

〈Abstract〉

Identification of potential technology convergence opportunities is crucial to drive innovation and growth in modern enterprises. In this study, we proposed a framework to explore technological convergence opportunities based on CPC code sequences from patents by utilizing the BERT model. We relied on the BERT architecture to train a new model using about 1.3 million patents registered at the Korean Intellectual Property Office, and achieved an accuracy of approximately 73% based on HitRate@10 metric. A case study using patents related to wearable technologies was conducted to demonstrate practicability and effectiveness of the proposed framework. The key contributions of this research include: (1) enabling in-depth analysis that takes into account the complex interactions between CPC codes and contextual variability; (2) enabling the exploration of diverse technology convergence scenarios beyond simple sequential patterns. This study is one of the first studies to apply the BERT model for exploring technology convergence opportunities, and is expected to contribute to the establishment of technology innovation and R&D strategies by providing a more accurate and practical tool for enhancing the speed and efficiency of technology opportunity-related decision-making processes.

Keywords : Technology Convergence, BERT, Patent Analysis, CPC, Wearables

1 제1저자, 박사과정, 경상국립대학교 대학원 기술경영학과
E-mail: skawk337@gnu.ac.kr
2* 교신저자, 조교수, 경상국립대학교 대학원 기술경영학과
E-mail: chsong01@gnu.ac.kr

1 First author, Graduate Student (Ph.D. program), Gyeongsang National University, Department of Management of Technology
2* Corresponding author, Assistant Professor, Gyeongsang National University, Department of Management of Technology

1. 서 론

기술의 급속한 진보로 인해 산업 환경의 복잡성, 가변성, 그리고 혁신성이 증대되고 있다. 이러한 동태적 환경에서 기업들은 지속적인 혁신과 적응의 필요성에 직면하게 되었으며, 이는 생존을 위한 경쟁을 더욱 심화시키고 있다. 이와 같은 경쟁의 심화는 기업이 기술적 우위 확보와 시장 지배력 강화를 위해 끊임없이 새로운 전략을 모색하고 실행하도록 하는 주요 동인으로 작용하고 있다. 이러한 맥락에서 기술융합은 기업의 혁신역량을 제고하고 경쟁우위를 창출하는 핵심 전략으로 부상하였다[1]. 기술 및 산업 간 패러다임의 전환을 이끌어온 기술융합은 산업구조와 시장환경뿐만 아니라 사회에서의 소통방식과 개인 간 연결방식을 변화시키는 핵심적인 역할을 해왔다[2]. 따라서, 융합에 기반한 기술혁신 역량은 기업들이 시장에서 차별화된 가치를 확보하고 변화하는 환경에 적응하며 지속적인 경쟁우위를 유지하기 위한 필수요소이다[3]. 특히, 새로운 비즈니스 모델을 창출하고 성장시키는 과정에서 내재된 위험을 해소하고 시장 경쟁력확보를 위한 잠재적인 기술융합기회 발굴에 관한 관심이 높아지고 있다. 기업들은 혁신적인 기술을 탐색하고 이를 효과적으로 상용화하여 시장에서의 입지를 강화하고 장기적으로 지속가능한 성장을 이루기 위한 전략적 의사결정을 지원하는 데 기술융합기회 발굴의 중요성을 강조하고 있다[4]. 하지만, 상당수의 중소기업은 기술혁신 전략을 통한 지속적인 성장과 변화 추구에 어려움을 겪고 있다. 이는 기술 기회 탐색에 대한 정보 부족이나 미래 사업화를 위한 역량 부족에 따른 것으로, 새로운 기술의 도입과 활용을 통한 사업 다각화와 성장 실현에 장애요인으로 작용한다[5]. 따라서 기업들은 기술융합기회를 체계적으로 탐색하고 이를 실제 사업화로 연계할 수 있는 역량을 강화할 필요가 있다.

무엇보다 전략적 의사결정을 내리기 위해서는 기술이 지닌 잠재성을 면밀히 평가하고, 미래에 다양한 가능성을 내포하고 있는 기술에 대해 심도 있는 분석 및 이해가 요구된다[6].

그동안 기술변화에 효과적으로 대응하고 기술기회를 선점하기 위해 체계적인 기술기회 발굴 프로세스 수립에 관한 연구가 활발히 진행되었으며[7], 선행연구에서는 기술기회 발굴을 위해 정성적 방법과 정량적 방법을 적용하였다. 델파이 분석과 같은 정성적 기술 예측 방법은 설문 대상자의 경험과 지식을 기반으로 기술 기회를 도출하지만, 이는 연구자의 주관성에 영향을 받는 한계를 지니고 있다. 반면, 정량적 분석 방법은 데이터 기반의 객관성을 확보할 수 있으나, 기술 간 복잡한 상호작용을 충분히 반영하지 못하는 단점이 존재한다. 기술융합기회 분석을 위해 CPC 코드의 동시출현(co-occurrence) 정도를 네트워크 분석에 접목하는 방법이 사용되었지만, 이는 빈도수나 네트워크의 구조적 특성만을 활용한다. 이러한 한계를 극복하기 위해 기술정보를 포함하는 특허와 같은 객관적인 자료를 기반으로 하는 새로운 방법론이 개발 및 적용되었다[8,9]. 그중 박진우·송지훈(2023)은 특허 분류코드에 Word2Vec 알고리즘을 적용한 후, 분류 코드 간 유사도를 산출해 웨어러블 기술영역에서의 잠재적 기술융합 기회를 추천하는 분석적 프레임워크를 제시하였다[10]. 해당 연구는 특허 코드의 시퀀스에 Word2Vec 알고리즘을 적용하고, 정적 임베딩 생성과 유사도 측정을 통해 기술융합기회를 추천한 데 의의가 있지만, 모델의 효용성에 대한 검증과정이 부재해, 실무적 활용도가 제한적이었다.

본 연구에서는 이러한 문제 인식을 기반으로 언급한 선행연구를 확장하여 BERT(Bidirectional Encoder Representations from Transformers)

모델을 활용한 기술융합기회 탐색 프레임워크를 제안한다. 이는 동시 출현하는 특허 코드의 쌍을 기준으로 기술융합기회를 특정하는 방법과는 근본적으로 차별화된다. BERT를 활용하는 모델은 특허 코드의 빈도나 통계적 특성보다는, 특허 코드 시퀀스가 지니는 기술맥락(technology context)의 이해를 기반으로 기술융합기회를 더 정밀하게 예측하게 해준다. BERT는 Transformer의 양방향 인코더를 심층적으로 쌓은 구조의 언어모델로 자연어 처리 분야에서 뛰어난 성능을 나타냈던 범용 모델이다[11]. BERT는 사전학습의 목적함수로 MLM(Masked Language Model)과 NSP(Next Sentence Prediction)를 사용한다. MLM은 입력된 문장의 토큰 중 일부를 [MASK] 토큰으로 변환하고 주변의 문맥 정보를 활용하여 마스킹 처리된 토큰을 예측하도록 하는 방법이다[12]. 마스킹 위치의 선정은 독립적으로 가능하기에, 단순 순차 패턴에 의한 기술융합기회가 아닌 문맥을 고려한 기술 간 연관성 분석이 가능하다. 본 연구에서는 특허의 CPC 코드 데이터를 활용해, BERT 구조와 MLM 방식을 적용함으로써 기술융합기회를 탐색하고자 한다. 이를 위해 대량의 CPC 시퀀스를 입력 데이터로 활용해 모델을 새롭게 학습시키고자 한다. 제안된 연구 프레임워크는 검증 과정을 거쳐 실용성을 입증하고자 하며, 도출된 결과는 향후 기술혁신 및 R&D 전략 수립에 기여할 것으로 기대된다. 특히, 기술기획 관련 의사결정이 더 신속히 이루어지는 데 일조할 수 있을 것으로 본다.

본 논문은 다음과 같이 구성된다. 2장에서는 데이터, 연구 프레임워크 및 실증적 연구방법에 관해 기술하며, 3장은 기술융합기회 탐색에 대한 분석 결과를 사례 기반으로 설명한다. 4장에서는 논문의 주요 연구 결과에 대해 토의하고, 한계점 및 향후 연구 방향을 제시한다.

2. 데이터 및 연구 방법론

본 연구는 Fig. 1에 제시된 연구 프레임워크를 기반으로 분석을 수행하였다. 연구 프레임워크는 총 4단계로 구성되며, 첫 번째 단계에서는 특허 데이터 수집 및 이에 대한 전처리 작업을 수행한다. 두 번째 단계에서는 특허의 CPC 특허분류 코드 데이터를 활용해 BERT 아키텍처를 기반으로 모델을 새롭게 학습시킨다. 세 번째 단계에서는 학습이 완료된 모델에 대한 성능검증을 수행한다. 마지막 단계에서는 학습된 신규 모델을 바탕으로 기술융합기회를 도출한다.

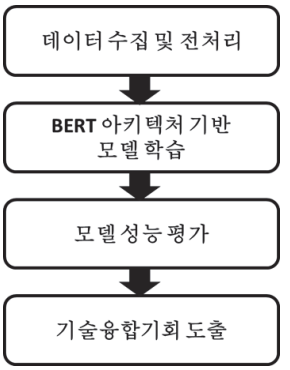

Fig. 1 Outline of the research framework

2.1 데이터 수집 및 전처리

본 연구는 국내 특허청에 등록된 특허 데이터만을 분석에 활용한다. 특허 데이터는 기술혁신 연구에서 널리 사용되는 기술 지식 생산의 핵심 지표이며, 분야별 기술 격차, 기술 트렌드, 미개척 혁신 영역을 식별해 연구자와 R&D 기획 담당자를 새로운 기술개발 기회로 안내하는 유용한 도구 역할을 담당한다. BERT 모델의 학습에는 대량의 시퀀스(sequence) 데이터가 요구되며 최신 기술 동향을 반영하고자 수집한 특허의 범위를 2010년

부터 2022년까지 출원된 데이터 중 등록된 데이터로 한정하였다. 특정 기술 분야에 대한 제한 없이 특허청에 출원 후 등록된 모든 데이터를 수집하였다. 본 연구에서 지칭하는 시퀀스 데이터는 출원된 특허에 할당된 CPC 코드 리스트를 의미한다. CPC는 발명의 특성에 따라 부여되는 국제적 특허 분류체계로 특허를 기술적 내용에 따라 검색 및 분류하는 목적으로 사용된다. 따라서, CPC 코드가 누락된 데이터 행의 경우 제거하는 과정을 거쳤다. 그 결과 총 1,345,240건의 데이터를 획득하였다. 데이터 수집은 특허 분석 서비스를 제공하는 윈텔립스 웹사이트를 통해 2024년 1월 수행되었다. CPC 코드는 계층적 구조로 구성되어 있으며, 각 코드 레벨에 따라 기술 특성을 더욱 세부적으로 정의하고 다양한 응용 분야를 나타낸다. 본 연구에서는 모델의 전반적인 복잡도를 고려해 메인그룹 레벨의 코드만을 학습에 활용하였다.

2.2 BERT 모델 사전학습

BERT는 본래 자연어 처리를 위한 사전학습 언어모델로 레이블이 되어있지 않은 텍스트 데이터를 사용하여 양방향으로 문맥을 학습하는 트랜스포머 인코더 구조를 기반으로 한다[11]. BERT는 word2vec과 달리 동일한 단어라도 문맥에 따라 다른 표현을 생성함에 따라 다의어 처리에 강점을 지녔으며, 질의응답, 감성 분석, 개체명 인식과 같은 다양한 자연어 처리 작업에 적용이 가능한 범용 모델이다. BERT는 MLM과 NSP를 통해 학습되며, MLM은 입력 시퀀스의 일부(15%)를 무작위로 마스킹하고 모델이 마스킹된 단어를 예측하는 방식을 의미한다. 즉, 시퀀스의 어느 위치에서든 독립적인 예측이 가능하다. NSP는 두 문장 간의 연속성 여부를 예측하는 작업으로 문장 간의 논리적 흐름을 학습한다. 따라서, BERT 모델은 단어뿐만 아니라 문장 수준에서 언어의 흐름을 더 깊게 이해하고 이를 바탕으로 다양한 자연어 처리에 있어 우수한 성능을 나타내었다. Binanchi et al.(2021)은 이러한 BERT 모델을 전자상거래(e-commerce) 분야에 접목한 연구를 발표하였다. 해당 연구는 Prod2BERT 모델을 제안하였고, 이는 BERT 기반의 문맥화된(contextualized) 제품 임베딩 모델을 지칭한다[13]. 모델의 입력 시퀀스로는 “자연어”가 아닌 쇼핑 세션(“online session log”)을 활용하며, 세션을 하나의 "문장"으로 세션 내의 제품들을 "단어"로 취급하여 모델의 학습에 활용한다. 모델의 학습에 있어 각 토큰은 제품 ID로 대체되며, Prod2BERT는 BERT 모델의 아키텍처를 모방해 쇼핑 세션의 문맥(context)에 따라 제품의 벡터표현(vector representation)을 학습하게 한다.

본 연구에서는 앞선 선행연구로부터 영감을 얻어 특허의 CPC 코드 데이터를 활용해 MLM 방식으로 BERT 아키텍처를 활용해 새로운 모델을 처음부터 학습시켰다. 즉, BERT 모델의 아키텍처를 기반으로 CPC 코드 데이터에 맞게 조정된 모델을 생성한다. 학습에 있어 유니크한 CPC 코드로 구성된 vocabulary를 생성하였으며, BERT 모델의 특성을 고려해 다음과 같은 특수 토큰 [PAD], [UNK], [CLS], [SEP], [MSK]가 추가되었다. [PAD]는 padding 토큰으로 문장의 길이를 일정하게 조정하기 위해 사용되며, [CLS]는 classification 토큰으로, 시퀀스의 시작을 나타내는 용도로 활용된다. 예를 들어, [H03F0001, H03F0003, H03F0003, H03H0007]의 오리지널 시퀀스의 경우 CPC 시퀀스의 길이를 10으로 설정할 때 다음과 같은 형태로 전처리되어 학습에 투입된다. ['[CLS]', 'H03F0001', 'H03F0003', 'H03F0003', 'H03H0007', '[SEP]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]']. MLM 학습 과정을 통해 CPC 코드의 문맥적 의미

가 상세히 반영된 임베딩이 생성되며, 이는 이종 기술 간의 결합정도를 효과적으로 표현한다. 이렇게 생성된 임베딩은 기술융합기회 발굴과 관련된 다운스트림 작업에 적용되어, 보다 정확하고 효율적인 분석을 가능하게 한다. 학습에 있어 CPC 코드의 최대 길이를 조정해가며 학습이 이루어졌다. 이는 특허의 종류에 따라 동시출현하는 CPC 길이의 편차가 큼을 반영하기 위함이다.

2.3 모델 성능 평가

본 연구에서는 신규 학습한 모델의 성능을 정확도 기준으로 측정하였다. 다만, 새롭게 학습한 BERT 기반 모델 자체의 정확도 기준이 아닌 기술융합기회 탐색의 목적에 따른 Top-k 방식을 따른다. Top-k는 추천시스템 연구에서 활용되는 평가지표로, k는 추천된 아이템의 수를 지칭한다. k개의 아이템은 모델이 사용자가 선호할 가능성이 가장 높은 순서로 정렬한 상위 k개의 리스트를 나타낸다. 정확도는 HitRate@k 지표로 측정되며, 이는 실제 채택된 CPC 코드가 추천된 Top-k 리스트에 포함된 경우 hit로 판정한다. 즉, CPC 코드 리스트에서 무작위로 하나의 코드를 제외한 후, 이에 대한 k개의 추천을 받은 리스트에 앞서 제외한 코드가 포함되면 hit로 간주한다. 모델의 성능은 이러한 과정의 반복을 통해 평균 hit 비율을 계산함으로써 최종적으로 평가된다. hit 비율은 일반적으로 k에 따라 증가하는 데, 이는 추천된 항목 수가 많을수록 관련 아이템을 발견할 확률이 높아짐을 시사한다. 본 연구에서 선택한 HitRate@k 지표는 기술융합기회 탐색이라는 연구 목적에 부합한다. 실제 기술 융합 과정에서는 다양한 기술 조합을 고려하게 되는데, Top-k 방식은 이러한 실제 상황을 잘 반영한다. 또한, HitRate@k는 모델의 추천 정확도를 다양한 k 값에 대해 평가할 수 있어, 모델의 성능을 더욱 포괄적으로 이해할 수 있게 해주며, 계산이 비교적 간단하면서도 모델의 성능을 직관적으로 이해할 수 있게 해준다.

2.4 기술융합기회 탐색

본 연구에서는 학습된 모델 중 가장 뛰어난 성능을 보인 모델을 기반으로, 특정 기술 도메인 분야의 CPC 데이터를 활용해 잠재적 기술융합기회에 대한 탐색을 진행한다. 본 연구의 모델은 위치 독립적 예측이 가능하기에, 사례기업의 데이터를 중심으로 추천되는 기술 코드에 대한 분석과 이로부터 도출할 수 있는 시사점에 대해 논한다.

3. 연구 결과

3.1 분석 데이터 개요

Fig. 2는 2010년부터 2022년 출원된 특허 중 등록된 특허의 변화 추이를 나타낸다. 통상적으로 특허의 등록까지는 출원 이후 약 2년 정도의 시간이 소요되기에, 2010년 출원 직후 등록된 수는 다소 낮게 나타난다. 특허 등록 추세는 2017년

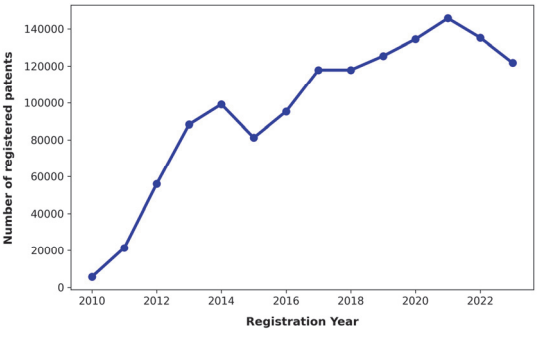

Fig. 2 Number of registered patents per year

이후 연 10만 건이 넘게 유지되는 것으로 나타났으며, 등록되는 특허의 수가 꾸준히 증가하고 있음을 확인할 수 있었다.

모델의 학습에는 특허에 할당된 CPC 코드 시퀀스가 활용된다. CPC 시퀀스의 평균 길이는 4.80, 표준편차는 2.98, 최대 길이는 57로 나타났다. 이는 입력으로 활용되는 시퀀스 대부분이 비교적 짧은 길이를 가지고 있음을 시사하며, 평균에 비해 상대적으로 큰 표준편차는 데이터의 분산이 큼을 시사한다. 최대 길이 57은 극단적으로 긴 샘플이 존재함을 의미하며, 이러한 이상치는 데이터의 특성을 왜곡시킬 수 있다. 따라서, 시퀀스의 평균 길이와 표준편차를 고려한 데이터의 입력 크기 결정이 요구된다.

Fig. 3은 전체 데이터 중 CPC 코드 상위 20개의 등장 빈도를 정리해 나타낸다. CPC는 계층적 구조를 지니기에, 바라보는 관점에 따라 세분화된 기술 수준에서의 분석을 가능하게 한다. 본 연구에서는 CPC의 메인그룹 레벨에서 분석을 수행하였다. 특정 기술 도메인에 국한되지 않은 전체 특허 데이터를 기준으로 분석을 수행하였기에 뚜렷한 경향성은 관찰되지 않았다. 하지만, H01L-0021, G06F-0003, G06Q-0050, H01M-0050, H01M-0010 순으로 코드의 등장 빈도가 높게 나타났으며, 이들은 모두 대한민국의 핵심 산업인 “반도체 및 관련 공정”, “컴퓨터 입출력 장치 및 인터페이스”, “비즈니스 모델”(비즈니스 관리 시스템) 및 “배터리” 분야와 밀접한 기술들을 나타낸다. 이러한 결과는 4차 산업혁명, 사물인터넷, 인공지능, 전기차 등 현재의 주요 기술 트렌드와 산업 동향이 반영된 결과로 볼 수 있다[10]. 반면, A61B-0005와 A61K-0008은 각 “의료 진단 및 측정 분야”와 “화장품(cosmetics)”과 관련된 기술을 나타내는 특허 분류 코드로 헬스케어와 뷰티 산업의 영역을 나타낸다. 최근 두 분야의 경계가 모호해지고 융합됨에 따라 코스메슈티컬(cosmeceutical)과 같은 새로운 제품 카테고리의 개발로 이어졌으며, 이는 단순한 기술융합을 넘어 새로운 시장 기회를 창출하는 현상으로 이어지는 변화를 수반한다[14].

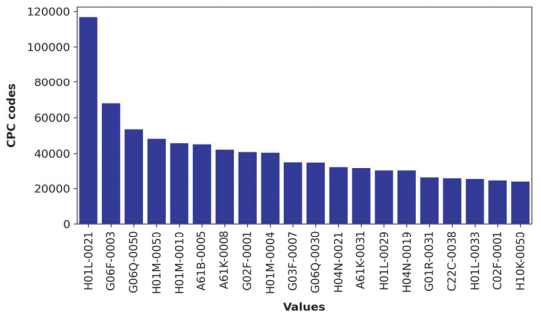

Fig. 3 Top 20 CPC code distribution

3.2 모델의 성능

본 연구에서는 CPC 시퀀스의 최대 길이를 조정하며 모델을 학습하는 방식을 채택하였다. 이는 입력 시퀀스 길이에 따른 모델의 성능변화를 측정하기 위함이다. 이러한 접근 방식은 모델의 계산 효율성을 향상하는데 용이할 뿐만 아니라 모델의 과적합(overfitting) 위험을 감소시키는 데 기여할 수 있다. 따라서, 본 연구에서는 CPC의 최대 길이를 10, 15, 20, 25 순으로 변화해가며 BERT 아키텍처 기반의 신규 모델을 학습시켰다. 모델 학습에는 전체 데이터 중 2010년부터 2019년 사이 출원 및 등록된 데이터가 활용되었다. 이 중 90%를 학습 데이터(training data), 학습 데이터의 20%를 검증 데이터(validation data)로, 그리고 10%를 평가 데이터(test data)로 활용하였다. 2020년에서 2022년 사이 출원 및 등록된 데이터는 HitRate@k 도출을 위해 사용되었다. 모델의 학습에는 TensorFlow 2.10.1 버전과 transformer 버전 4.24.0이 활용되었다.

Table 1. Vocabulary size and model performance based on changing length of CPC sequence

CPC 시퀀스 길이	Vocabulary size	Validation accuracy	Test accuracy
10	10059	0.4069	0.4226
15	10075	0.4075	0.4257
20	10079	0.4074	0.4209
25	10083	0.4079	0.4210

Table 1은 CPC 시퀀스의 최대 길이 조정에 따른 vocabulary 사이즈와 모델의 성능을 나타낸다. Vocabulary는 언어모델이 처리할 수 있는 고유한 토큰들의 집합을 나타내며, 토큰화된 입력 시퀀스는 vocabulary 내에서 고유한 인덱스로 맵핑되어 모델에 입력된다. Vocabulary의 사이즈에 따라 모델의 일반화 성능이 향상될 수 있지만, 동시에 학습 시간과 비용("계산 효율성")은 저하될 수 있다. Vocabulary 사이즈는 시퀀스 길이가 증가함에 따라 소폭 증가했지만, 변화의 폭은 미비하다고 볼 수 있다. 테스트 데이터 기준 모델의 성능은 전반적으로 42%에 해당했으며, 시퀀스 길이에 따른 편차는 크지 않음을 확인할 수 있었다.

Table 2는 HitRate@k를 활용한 모델 간 성능 비교를 나타낸다. K의 개수는 5와 10개를 기준으로 테스트를 수행하였다. 그 결과 k가 증가하는 경우, HitRate 또한 개선됨을 확인할 수 있었으며, CPC 시퀀스의 최대 길이가 15인 경우 가장 높은 정확도를 보였다. 이는 반드시 더 긴 시퀀스의 활용이 성능을 보장하지 않음을 의미한다. 즉, 일반적인 자연어가 아닌 CPC 데이터의 특수성과 정보

Table 2. Model comparison using HitRate@k

CPC 시퀀스 길이	HitRate@5	HitRate@10
10	57.2%	64.9%
15	65.1%	72.8%
20	55.5%	63.1%
25	50.8%	58.1%

분포에 의하면 시퀀스 길이가 15개인 경우 최적 성능을 나타내는 것으로 분석되었다.

3.3 사례분석을 통한 기술융합기회 탐색

본 연구에서는 "웨어러블" 분야의 기술융합기회 탐색을 사례로 선정해 분석을 진행하였다. 웨어러블은 선행연구를 통해 이미 다양한 기술 분야와의 융합으로 새로운 가치를 창출하고 있는 기술 영역으로 알려져 있다[10]. 이러한 배경에서, 웨어러블 내 세부기술 분야와 융합 가능성이 높은 분야에 대한 체계적인 탐색은 의미 있는 연구 주제로 판단되었다. 이를 위해 웨어러블 기술 관련 국내 특허 데이터를 추가로 수집하였고, 그중 CPC 시퀀스의 길이가 15 이하인 예제 2개를 선별해 탐색하였다. 웨어러블 기술 관련 특허는 선행연구[10]에 언급된 검색식을 참조해 획득하였다.

첫 번째 특허의 CPC 시퀀스는 다음과 같이 정의된다: "G16H-0050, G06Q-0050, H04W-0004, A61B-0007, A61B-0005, A61B-0005, A61B-0005, G06N-0020". 새로운 기술융합기회 탐색에 있어 연구자는 원하는 위치의 CPC 코드 마스킹 또는 마지막 코드에 이어 등장 가능한 CPC 코드를 예측하는 방식을 활용할 수 있다. 본 사례에서는 전자를 택하였다. G16H-0050은 "의료 진단, 의료 시뮬레이션 또는 의료 데이터 마이닝에 특히 적합한 ICT; 유행병 또는 전염병의 탐지, 모니터링 또는 모델링에 특히 적합한 ICT", A61B-0007은 "청진기기"를 의미한다. 위 시퀀스에서 각 G16H-0050과 A61B-0007을 마스킹 처리 후 예측을 진행한 결과는 다음과 같다. 전자의 경우 ['G06Q-0050', 'A61B-0005', 'G06Q-0010', 'G16H-0020', 'H04W-0004'], 후자의 경우 ['H04W-0004', 'A61B-0005', 'H04W-0088', 'G06Q-0050', 'H04M-0001']의 코드들이 추천되었다. 코드가 서로 다른 위치에 놓여있고 특성이 다름에도 불구하고 중복되는

CPC 코드가 추천되었는데, 이는 G16H-0050과 A61B-0007이 서로 밀접한 연관성을 지님을 시사한다. 특히 의료 진단 및 측정 기술(A61B-0005)이 G06Q-0050과 관련성이 높음을 시사한다. 두 기술 코드의 조합은 웨어러블 헬스케어 기기 및 시스템 개발에 활용될 수 있는 핵심 기술 영역으로 작용할 수 있음을 나타낸다.

두 번째 특허의 시퀀스는 다음과 같이 표현된다: “A01K-0029, A01K-0011, G06Q-0050”. A01K-0029는 “기타 가축 용구”, A01K-0011은 “동물의 표식붙이기”, G06Q-0050은 “특정 사업 부문의 사업 프로세스 구현에 특히 적합한 정보통신기술[ICT]”을 지칭한다. 위 시퀀스에서 각 A01K-0011과 G06Q-0050을 마스킹 처리 후 예측을 진행한 결과는 다음과 같다 전자의 경우 ['G06Q-0050', 'H04W-0004', 'G06Q-0020', 'G06Q-0010', 'G06Q-0030'], 후자의 경우 ['G06K-0019', 'G06Q-0050', 'H04W-0004', 'H04N0-007', 'G06Q-0020']의 코드들이 추천되었다. 공통으로 G06Q-0050, G06Q-0020 및 H04W-0004가 추천되었으며, 해당 특허(웨어러블 장치를 통한 반려동물에 대한 모니터링 기술)에 무선네트워크 설비 서비스(H04W-0004)를 추가하거나 지불 방식 프로토콜(G06Q-0020)을 결합하는 방식으로 반려동물 모니터링 분야 기술발전 방향에 대한 제시가 가능하다.

4. 결 론

본 연구에서는 언어모델인 BERT를 자연어 처리 태스크(task)가 아닌 기술융합기회 탐색을 위한 용도로 접목시켜, 특허의 CPC 시퀀스 데이터를 입력값으로 한 모델을 생성하였다. 이를 위해 국내 특허청에 출원 및 등록된 특허 약 백삼십만 건을 활용해 모델에 대한 학습 및 평가를 진행하였다. 그 결과 HitRate@10을 기준으로 약 73%의 정확도를 달성할 수 있었다. 본 연구의 주요 의의는 BERT의 맥락화(contextualized) 임베딩 방식을 특허의 CPC 코드 시퀀스에 최초로 적용한 데 있다. 기존 연구에서 co-classification을 기반으로 기술융합 패턴을 분석하는 방법이 주로 사용됐다면, 본 연구에서는 CPC 시퀀스 데이터를 입력값으로 하는 언어모델 제시를 통해 기술융합기회를 탐색하는 새로운 접근법을 제시하였다. 이를 통해 함께 등장하는 CPC 코드 쌍의 빈도나 단순한 코드 간 유사성 분석을 넘어, CPC 코드 간의 복잡한 상호작용과 맥락적 변동성을 고려한 심층적인 분석이 가능해졌다.

본 연구에서 생성한 모델은 연구자가 희망하는 유동적인 CPC 조합을 입력값으로 활용해, 위치 및 길이 독립적으로 조화를 이룰 수 있는 CPC 시퀀스에 대한 예측을 수행할 수 있다. 이러한 특성은 기존의 순차적 모델들과 차별화되며, 다양한 기술융합기회를 탐색하는 데 유용하게 활용될 수 있다. 무엇보다 특정 기술 분야의 전문가가 자신의 전문 분야의 기술과 융합 가능성이 존재하는 타 분야의 기술 코드를 입력하여, 새로운 융합 기술 영역을 발견할 수 있다.

본 연구는 기술융합 연구에 있어 새로운 방법론을 제시하였지만, 다음과 같은 한계점과 향후 연구 방향을 고려할 수 있다. 현재 학습시킨 모델의 정확도는 실무적으로 적용되기에는 다소 낮다고 볼 수 있으며, 시퀀스 길이 외 파라미터에 대한 조정을 통해 모델의 성능을 개선하는 시도가 요구된다. 또한, 국내 데이터만이 이번 연구에 활용되었다는 점에서, 추후 미국이나 유럽에 등록된 특허 데이터를 활용해 추천되는 국가 간 기술융합기회의 차이를 비교 분석하는 후속 연구를 수행할 수 있다. 향후 연구에서는 이러한 접근법을 특정 기술 도메인에 특화된 기술융합기회 탐색 프레임

워크 개발에 적용할 수 있다. 마지막으로, 탐색된 기술융합기회의 결과를 실제 특허 문서의 내용 분석을 통해 보다 심층적으로 검증하는 것도 의미 있을 것이다.

사 사

본 논문은 산업통상자원부의 '융합기술사업화 확산형 전문인력 양성사업'의 지원을 받아 수행된 논문임.

참고문헌

[1] Hwang, J., Kim, K. H., Hwang, J. G., Jun, S., Yu, J., Lee, C. Technological opportunity analysis: Assistive technology for blind and visually impaired people, Sustainability, 12, 20, 8689, p.1-17, (2020).
[2] Song, C. H., Elvers, D., Leker, J. Anticipation of converging technology areas — A refined approach for the identification of attractive fields of innovation, Technological Forecasting and Social Change, 116, p.98-115, (2017).
[3] 김성홍, 기술혁신역량이 개방형혁신활동 및 기술혁신성과에 미치는 영향, 경영컨설팅연구, 24, 1, p.131-145, (2024).
[4] Lee, C., Kang, B., Shin, J. Novelty-focused patent mapping for technology opportunity analysis, Technological Forecasting and Social Change, 90, p.355-365, (2015).
[5] 김상욱, 최경현, 인수합병 거래 데이터를 활용한 헬스케어 기술기회탐색, KSBB Journal, 36, 1, p.1-15, (2021).
[6] 이지호, 정병기, 고남욱, 오승현, 윤장혁, 특허의 Problem-Solution 텍스트 마이닝을 활용한 기술경쟁정보 분석 방법, 지식재산연구, 13, 3, p.171-204, (2018).
[7] 서원철, 기술테마 분석을 통한 기술기회발굴 연구-3D 프린팅 기술 사례를 중심으로, 지식재산연구, 16, 2, 205-248, (2021).
[8] 박영진, 고남욱, 윤장혁, 보유특허 기반의 기술기회탐색을 위한 특허추천방법: 3D 프린팅 산업을 중심으로: 3D 프린팅 산업을 중심으로, 지식재산연구, 10, 1, p.169-200, (2015).
[9] 유경영, 송지훈, 특허정보를 활용한 디지털 트윈 기술 동향 분석 및 기술융합기회 발굴, 한국산업융합학회논문집, 26, 3, p.471-481, (2023).
[10] 박진우, 송지훈, Word2vec 기반의 기술융합기회 발굴 연구: 웨어러블 기술사례를 중심으로, 한국산업융합학회논문집, 26, 5, p.833-844, (2023).
[11] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint, arXiv:1810.04805, (2018).
[12] 고영수, 이수빈, 차민정, 김성덕, 이주희, 한지영, 송민, BERTopic 을 활용한 불면증 소셜 데이터 토픽 모델링 및 불면증 경향 문헌 딥러닝 자동분류 모델 구축, 정보관리학회지, 39, 2, p.111-129, (2022).
[13] Bianchi, F., Yu, B., Tagliabue, J. BERT goes shopping: Comparing distributional models for product representations, arXiv preprint, arXiv: 2012.09807, (2020).
[14] Park, S. H., Kwon, H. J. Customers' convergent recognition and satisfaction about cosmeceuticals, Journal of Digital Convergence, 15, 2, p.459-464, (2017).

(접수: 2024.07.03. 수정: 2024.07.30. 게재확정: 2024.08.06.)